**本科生毕业设计（论文）开题报告表**

|  |  |  |  |
| --- | --- | --- | --- |
| 毕业设计（论文）题目：基于模型融合的在线招聘欺诈检测 | | | |
| 题目类型 | **应用研究（☑）理论研究（）** | 题目来源 | **导师科研（☑）实习实践（）** |
| 指导教师 | **李岩** | 开题日期 | **2025年3月13日** |
| 学生姓名 | 周伟诺 | 专业年级 | 大数据管理与应用21-2班 |
| 1. ****立项依据****    1. 选题背景   近年来，互联网的快速发展深刻改变了人力资源市场的招聘模式。传统的线下招聘形式逐渐被新型的招聘方式所取代，网络招聘行业如雨后春笋般涌现，成为企业招聘和求职者求职的重要桥梁。  图表 1：2022年中国网络招聘行业市场规模  这种变革不仅体现在网络招聘平台的多样化上，还体现在技术手段的创新和应用上。随着云计算、大数据、人工智能（Artificial Intelligence，AI）等前沿技术的不断成熟，招聘形式也逐渐朝着智能化和数字化的方向发展。例如，云招聘的兴起，让企业能够通过线上平台发布招聘信息、筛选简历，并组织视频面试，从而降低了传统招聘过程中时间和地域的限制。此外，社交媒体招聘也逐渐成为一种新的趋势，通过社交平台如小红书、LinkedIn和脉脉，以及短视频平台如抖音、快手和微信视频号等等，企业不仅能广泛宣传招聘信息，还能直接与潜在候选人互动，从而提升招聘效果。同时，人工智能技术的应用进一步提升了招聘的智能化水平，基于AI的招聘系统可以实现简历的自动筛选、候选人匹配以及智能推荐，有效减少了人力资源部门的重复性工作，提高了整体招聘效率。  图表 2：招聘平台及招聘形式多样化  然而，网络招聘行业在迅速发展的同时，虚假招聘信息也逐渐泛滥，严重影响了网络招聘市场的健康发展。根据Greenhouse的研究，约18%到22%的在线招聘广告实际上并没有招聘意图，这些被称为“幽灵职位”的广告不仅对求职者造成困扰，还导致了对潜在雇主的不信任感。虚假招聘信息通常通过虚构职位、夸大薪资待遇、伪装成知名企业等手段诱导求职者上当，最终通过获取个人信息、收取押金或培训费用等方式牟取非法利益。这些虚假招聘信息形式多样，不仅让求职者的时间和精力付之东流，还可能导致其个人隐私泄露、经济损失，甚至身陷诈骗陷阱。虚假招聘信息的出现，给社会带来了诸多负面影响。一方面，求职者在面对大量虚假招聘广告时，常常无法有效辨别真假，导致个人信息泄露、金钱被骗，甚至一些人因高薪诱惑而误入非法传销组织。另一方面，虚假招聘信息的泛滥严重影响了求职者对网络招聘平台的信任，阻碍了招聘生态的健康发展。企业则面临虚假招聘广告冒用其名义发布信息的风险，这不仅损害了企业形象，还可能导致求职者对该企业产生误解，进而影响其品牌声誉。  目前，尽管部分网络招聘平台已建立了举报和人工审核机制，但由于虚假招聘信息的数量庞大、发布渠道复杂，传统的人工审核方法不仅效率低下，而且难以覆盖所有的虚假信息。尤其随着招聘广告形式日益多样化，人工审核的局限性愈发显现。例如，仅用基于规则的模型，难以应对表述越来越灵活的虚假信息，无法在实际应用中很好地满足快速、精准识别虚假招聘信息的需求。因此，如何通过技术手段实现对虚假招聘信息的快速识别，已成为招聘行业亟待解决的难题。   * 1. **研究问题与意义**      1. 研究问题  1. 如何识别本质上具有欺诈性的职位描述的关键特征？ 2. 如何创建和优化一个使用文本数据和元特征的分类模型，并预测哪些职位是虚假的或真实的？ 3. 如何评估和对比不同分类模型在虚假招聘信息识别中的性能？应考虑哪些模型评价指标来判断融合模型相较于单一模型的优势和不足？ 4. 虚假招聘信息的识别方法如何与现有的招聘平台系统进行有效整合，确保在实际应用场景中能够流畅、高效地运行？    * 1. 研究意义   （1）       保护求职者的利益：求职者是虚假招聘信息的主要受害者。虚假招聘信息通常通过夸大的薪资待遇、虚构的职位要求等手段诱骗求职者，造成时间、金钱和个人隐私的损失。  （2）       推动招聘行业的健康发展：虚假招聘信息不仅影响求职者和平台，也可能扰乱招聘市场的正常秩序。  （3）       助力监管与平台管理：监管部门在打击虚假招聘乱象时，需要有效的技术手段来辅助判断信息的真实性。模型融合的识别方法可为监管部门提供可靠的数据支撑和判断依据，使其能更精准地发现并查处发布虚假招聘信息的主体，增强监管的有效性，维护良好的市场监管秩序。   * 1. **文献综述**   随着虚假招聘广告的泛滥，如何有效地识别虚假招聘信息成为了研究的一个重要课题。近年来，学者针对中文在线招聘欺诈的研究少之又少，由于没有公开的中文在线欺诈检测数据集，因此相关的研究进展集中在半监督学习方向。国内学者提出了一种基于一致性训练的半监督虚假招聘广告检测模型（SSC）。该模型通过引入一致性正则项来提升模型的性能，具体方法是通过对所有数据应用一致性正则项，在标签数据较少的情况下仍能获得较好的识别效果。进一步地，SSC模型通过联合训练的方式整合有监督损失和无监督损失，最终优化模型的整体性能。  为了检测虚假招聘信息，Vidros 等人正式发布了第一个英文数据集爱琴海就业欺诈数据集（EMSCAD），包含17880条带多维标签的招聘数据集，并将传统的机器学习分类器应用于该数据集以检测在线招聘欺诈。他们进行了两类实验并比较了结果。第一个实验包括六个不同的分类器：Naive Bayes (NB)、Zero Rule (ZeroR)、One Rule (OneR)、Logistic Regression (LR)、J48 和 Random Forest (RF)。本实验的最佳分类器是 RF，精确度最高，达到 91.4%。第二个实验使用了经验规则集模型。LR、J48 和 RF 分类器为经验规则集建模提供了 90.6% 的精度。Akram等建立包含语言风格、公司信息真实性等12项评估指标的综合评价体系，为模型优化提供多维参照。基于EMSCAD数据集，学者们采用了多种机器学习方法进行虚假招聘广告的检测[4-11]。例如，早期研究借鉴垃圾邮件检测思路，采用支持向量机（SVM）、随机森林等算法。Dutta等通过词频统计和规则匹配实现初步检测，但局限于浅层特征。Alandjani比较了多种分类器性能，发现集成方法在中小规模数据上表现稳定。Shibly等对比发现决策森林在中小数据集上优于提升决策树，为模型选择提供了实证依据。Jagadeesh等创新性引入岗位薪资与资历匹配度指标，使随机森林模型的F1值提升9.2%。这类方法在可解释性方面优势明显，但依赖专家经验的特征工程制约了泛化能力,这类方法依赖人工特征工程，在复杂语义场景中存在瓶颈。基于自然语言处理技术，研究者开始探索深度学习模型。Bidirectional LSTM和CNN通过捕捉文本序列的深层语义关系，显著提升检测准确率。Kim创新性地结合层次聚类与深度神经网络，实现对欺诈模式的细粒度识别。深度学习在非结构化文本处理上展现优势，但面临标注数据稀缺的挑战。针对标注成本问题， Cheekati Srikanth通过Stacking集成策略融合XGBoost、LightGBM等基模型，在公开数据集上F1值达92.3%。这类混合方法在保持模型鲁棒性的同时，缓解了数据不足的局限。研究趋势从单一文本分析转向多特征融合。谢宁宁等构建包含52维统计特征与语义特征的混合向量，准确率提升11.6%  Chiraratanasopha引入行业知识特征（薪资异常值、联系方式规范性等等），有效识别传统NLP方法易忽略的欺诈模式。由于虚假招聘样本的天然稀缺性导致类别失衡，Afzal等采用SMOTE-Tomek联合采样法，将少数类样本扩充3倍，使模型召回率提高19%。Vo等设计代价敏感学习框架，通过动态调整损失函数权重平衡分类效果，在1:8的失衡数据下保持85%的F1-score。Rofik等应用GridSearchCV优化SVM核函数和梯度提升树深度，将检测速度提升40%。Naudé提出基于贝叶斯优化的自适应调参策略，在跨领域数据迁移场景下表现突出。Choudhury等将遗传算法引入特征选择，通过自适应权重调整使关键特征维度缩减40%而性能不变。Alghamdi等开发实时特征更新管道，动态跟踪欺诈模式演变，在持续学习场景下保持83%检测稳定性。这类方法有效应对了欺诈策略的动态变化Lal等开发ORFDetector系统，集成流式计算与滑动窗口机制，处理延迟<200ms。Reddy等设计基于Flask的Web应用，支持岗位URL实时验证，日均检测量超5000条。   1. 研究内容   针对提出的研究问题拟展开以下四个方面的研究内容：   1. 提取虚假招聘信息中的语义特征 2. 模型选择与模型融合策略 3. 模型融合性能评估与对比 4. 模型部署与发布    1. 提取虚假招聘信息的语义特征   首先对文本数据进行预处理，包括文本清理、去除停用词和词干化与词形还原等等，在数据预处理后，使用常见的特征提取方法（Bag-of-Words、TF-IDF、Word Embeddings）来提取相关的语义特征, 应用特征选择算法（如卡方检验、信息增益、L1正则化）去除冗余特征，或利用PCA等降维技术减少特征空间，提高计算效率，从提取的语义特征中，结合其他分类特征与数值型特征构建可以用于机器学习模型的特征。   * 1. 模型选择与模型融合策略   本研究考虑运用多种模型融合策略，包括加权法、投票法、stacking方法等方法，旨在降低模型过拟合风险，增强模型的泛化能力，进而提高对虚假信息的识别效果。加权法通过为不同模型的输出赋予特定权重，综合考虑各模型优势，以获得更准确的预测结果；投票法则基于多个模型的预测结果进行简单投票，选择得票最多的类别作为最终预测；stacking方法是将多个基础模型的输出作为新的特征输入到一个元模型中进行进一步学习和预测。在此基础上，本研究重点聚焦于如何根据不同模型在实际应用中的性能表现，动态调整融合策略中的权重参数。通过深入分析各模型在不同数据集和任务场景下的表现，利用智能算法或自适应机制，实现权重的动态优化。   * 1. 模型融合性能评估与对比   评估并对比不同模型融合策略在虚假招聘信息识别中的性能，找出最佳融合方案；其次可以通过交叉验证等方法，评估模型融合是否能有效避免过拟合，并且在未见数据上保持较好的泛化能力。   * 1. 模型部署及发布   基于云服务器搭建以Linux操作系统为基础的部署环境，安装好所需软件与依赖库，将训练好的模型进行打包、优化处理。随后运用 FastAPI 框架构建 RESTful架构的API服务，明确输入输出格式并加入错误处理与日志记录功能。接着将模型发布到生产环境，设置访问权限保障安全，确保模型稳定、高效运行，为实际应用提供有力支持。   * 1. 研究方法   本研究的研究方法主要有四种：   1. **实证分析法**：实证分析法是深入探究在线招聘信息检测问题的重要手段。首先，明确定义一系列与研究紧密相关的核心变量，通过定义职位描述详细程度、薪资合理性、公司信息完整性等核心变量构建多维评估框架，进而建立针对性模型以揭示变量间的关联性及其对信息真实性的影响机制，最终结合描述性统计与相关性分析，系统评估招聘信息的可信度，从而为招聘真实性判断提供量化依据。 2. **文本挖掘法**：鉴于招聘信息主要以文本形式呈现，文本挖掘法在本研究中发挥着关键作用。针对研究内容的特点和需求，首先对收集到的数据进行全面细致的清理，去除其中的噪声信息。这些噪声可能包括格式错误、乱码、无关的特殊字符以及重复冗余的内容等。通过数据清理，能够提高数据质量，为后续的特征提取提供纯净、有效的数据基础。在清理后的数据基础上，运用先进的自然语言处理技术，从招聘信息中深度提取与虚假信息识别紧密相关的语义特征。这些语义特征涵盖多个层面，例如词汇层面，通过分析招聘信息中特定词汇的出现频率、词向量等，挖掘出与虚假信息相关的高频词汇或具有特殊语义指向的词汇；句法层面，研究句子结构、语法规则以及词性搭配等方面的特征，识别出不符合正常招聘信息表达习惯的句式。通过文本挖掘法提取的语义特征，能够从文本的本质层面为虚假招聘信息的识别提供丰富、准确的线索，有助于提高识别模型的准确性和有效性。 3. **实验研究法**：构建多个不同类型的基础分类模型用于招聘信息检测，如逻辑回归模型、决策树模型、随机森林模型、支持向量机模型、神经网络模型等。针对不同模型设置不同的参数组合，开展多组实验，对比各模型在相同数据集上的性能表现，包括准确率、召回率、F1 值等指标。 4. **模型融合法**：单一模型在面对复杂多变的在线招聘信息时，往往存在一定的局限性，难以全面、准确地识别虚假招聘信息。为了克服这一问题，本研究采用模型融合法，旨在通过整合多个模型的优势，提高虚假招聘信息识别的准确性和鲁棒性。模型融合法的核心思想是将不同类型的基础模型在处理招聘信息时所捕捉到的信息和特征进行有机结合。这些基础模型由于各自的算法原理和学习方式不同，对招聘信息的理解和表征也有所差异。在模型融合过程中，采用多种融合策略。常见的融合策略包括投票法、均值法和Stacking法等。通过模型融合法，能够充分整合各个基础模型的优势信息，弥补单一模型的不足，有效提高虚假招聘信息识别的准确性和鲁棒性。同时，面对招聘信息数据的动态变化和噪声干扰，融合模型能够展现出更强的适应性和稳定性，为实际应用提供更可靠的虚假招聘信息识别解决方案。    1. 数据集描述   本研究的数据集来源于公开数据集[fake-job-postings](https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction)、[FakePostings](https://www.kaggle.com/datasets/srisaisuhassanisetty/fake-job-postings)和[58同城招聘数据集](https://github.com/freekatz/ORFD/blob/master/orfd/Core/dataset/origin.csv)，即本研究的数据集来源于三份数据集，分别为fake-job-postings、FakePostings和58同城招聘数据集。其中前两份数据集为英文数据集，58同城招聘数据集为中文数据集，为了便于区分，本研究将其分别命名为the first dataset、the second dataset和58同城招聘数据集，其中前两份数据集合并之后用于**训练模型**，58同城招聘数据集用于**检验模型**的通用性。   * + 1. The First Dataset   此数据集来自公开数据集[FakePostings](https://www.kaggle.com/datasets/srisaisuhassanisetty/fake-job-postings), 它包含了一系列在线招聘广告的文本数据，这份数据集包含10000条虚假招聘信息的精选列表。该数据集主要用于研究欺诈性招聘模式和诈骗策略，通过分析职位描述、公司信息、薪资待遇等特征，来判别广告的真实性。字段及描述如下表所示：  表格 1：第一份数据集的字段类型及描述   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 字段 | 有效值数量 | 缺失值 | 缺失值比率 | 数据类型 | 变量类型 | | title | 10000 | 0 | 0 | String | 文本型 | | description | 10000 | 0 | 0 | String | 文本型 | | requirements | 10000 | 0 | 0 | String | 文本型 | | company profile | 10000 | 0 | 0 | String | 文本型 | | location | 10000 | 0 | 0 | String | 分类型 | | salary range | 10000 | 0 | 0 | String | 离散定量型 | | employment type | 10000 | 0 | 0 | String | 分类型 | | industry | 10000 | 0 | 0 | String | 有序分类型 | | benefits | 10000 | 0 | 0 | String | 文本型 | | fraudulent | 10000 | 0 | 0 | int | 二元型 |  * + 1. The Second Dataset   此数据集来源于公开数据集EMSCAD，该数据集包含17880份在线招聘广告，其中866份是虚假的招聘信息，数据包括有关职位的文本信息和元信息，可以识别本质上具有欺诈性的职位描述的关键特征和特性，或创建分类模型来预测哪些职位描述是虚假的或真实的。属性及描述信息如下表所示：  表格 2：第二份数据集的字段类型及描述   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 字段 | 有效值 | 缺失值 | 缺失值比率 | 数据类型 | 变量类型 | | job id | 17880 | 0 | 0 | int | 离散型 | | title | 17880 | 0 | 0 | String | 文本型 | | location | 17534 | 346 | 0.0194 | String | 分类型 | | department | 6333 | 11537 | 0.6458 | String | 分类型 | | salary range | 2868 | 15012 | 0.8396 | String | 定量型 | | company profile | 14572 | 3308 | 0.1850 | String | 文本型 | | description | 17879 | 1 | 0.0001 | String | 文本型 | | requirements | 15184 | 2696 | 0.1508 | String | 文本型 | | benefits | 10668 | 7212 | 0.4034 | String | 文本型 | | telecommuting | 17880 | 0 | 0 | int | 二元型 | | has company logo | 17880 | 0 | 0 | int | 二元型 | | has questions | 17880 | 0 | 0 | int | 二元型 | | employment type | 14409 | 3471 | 0.1941 | String | 分类型 | | required experience | 10830 | 7050 | 0.3943 | String | 有序分类型 | | required education | 9775 | 8105 | 0.4533 | String | 有序分类型 | | industry | 12997 | 4903 | 0.2743 | String | 分类型 | | function | 11425 | 6455 | 0.3610 | String | 分类型 | | fraudulent | 17880 | 0 | 0 | int | 二元型 |  * + 1. 58同城招聘数据集   此数据集来源于[58同城招聘网站](https://github.com/freekatz/ORFD/blob/master/orfd/Core/dataset/origin.csv)，该数据集是包含了868条数据且带标签的中文虚假招聘数据集，其中434条为虚假招聘数据。数据集主要包括文本信息和元信息，主要属性及描述信息如下表所示：  表格 3：58同城招聘数据集的字段类型及描述   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 字段 | 有效值 | 缺失值 | 缺失值比例 | 数据类型 | 变量类型 | | jobTitle | 868 | 0 | 0 | String | 文本型 | | jobSalary | 868 | 0 | 0 | String | 定量型 | | jobWelfare | 847 | 21 | 0.024 | String | 文本型 | | jobRequirement | 868 | 0 | 0 | String | 文本型 | | jobDescription | 868 | 0 | 0 | String | 文本型 | | companyIntro | 868 | 0 | 0 | String | 文本型 | | jobCity | 868 | 0 | 0 | String | 分类型 | | Real/Fake | 868 | 0 | 0 | int | 离散型 |   考虑到当前许多模型在训练过程中所采用的数据具有一定局限性，为了充分验证模型在不同语言环境下的表现和适用性，因此这个数据集主要用于验证模型的通用性。   * + 1. 数据合并与清理   在合并和清理这两个英文数据集的过程中，首先需要确保每个字段的数据一致性，且在合并后不会丢失关键信息，以下是对数据字段进行处理后的描述。  表格 4：文本型变量描述信息   |  |  |  | | --- | --- | --- | | 属性 | 描述 | 数据类型 | | title | 职位标题 | String | | description | 职位描述 | String | | requirements | 职位要求 | String | | company profile | 公司简介 | String | | benefits | 公司福利 | String |   表格 5：分类型变量描述信息   |  |  |  |  | | --- | --- | --- | --- | | 属性 | 描述 | 数据类型 | 样本 | | location | 公司位置 | String | us | | employment type | 招聘类型 | String | fulltime | | industry | 所属行业 | String | health care | | department | 所属部门 | String | marketing | | required experience | 需要的经验 | String | internship | | required education | 需要的教育水平 | String | bachelor and above | | function | 岗位职能 | String | education |   表格 6：二元变量描述信息   |  |  |  |  | | --- | --- | --- | --- | | 属性 | 描述 | 数据类型 | 取值 | | telecommuting | 是否远程工作 | int | (0,1) | | has company logo | 是否有公司logo | int | (0,1) | | has questions | 是否有问题 | int | (0,1) | | is missing required experience | required experience是否缺失值 | int | (0,1) | | is missing employment type | employment type是否缺失值 | int | (0,1) | | is missing industry | industry是否缺失值 | int | (0,1) | | is missing required education | required education是否缺失值 | int | (0,1) | | is missing requirements | requirements是否缺失值 | int | (0,1) | | is missing department | department是否缺失值 | int | (0,1) | | is missing telecommuting | telecommuting是否缺失值 | int | (0,1) | | is missing has company logo | has company logo是否缺失值 | int | (0,1) | | is missing has questions | has questions是否缺失值 | int | (0,1) | | is missing company profile | company profile是否缺失值 | int | (0,1) | | fraudulent | 是否为欺诈 | int | (0,1) |  * 1. 研究技术路线图   图表 3：研究技术路线图   * 1. 论文章节安排  1. 绪论    1. 研究背景与意义       1. 在线招聘行业现状       2. 招聘欺诈问题的严重性       3. 研究的目的和意义    2. 文献综述       1. 招聘欺诈检测研究现状       2. 多模型融合方法研究现状       3. 总结与分析    3. 研究内容与方法       1. 研究内容概述       2. 研究方法与技术路线图    4. 论文组织结构安排 2. 相关理论与技术基础    1. 在线招聘欺诈相关概念       1. 在线招聘欺诈的定义与类型       2. 在线招聘欺诈的特征与识别难点    2. 自然语言处理技术       1. 自然语言处理技术的应用       2. 常用的自然语言处理工具    3. 特征工程       1. 特征工程的概念与重要性       2. 特征编码    4. 机器学习基础模型       1. 逻辑回归模型       2. 决策树模型       3. 随机森林模型       4. 支持向量机模型       5. 神经网络模型    5. 多模型融合方法       1. 投票法       2. 均值法       3. Stacking方法       4. Blending方法    6. 本章小结 3. 在线招聘数据收集与预处理    1. 数据收集       1. 数据来源       2. 公开数据集介绍    2. 数据集描述       1. 数据集规模与特征       2. 数据的标签信息    3. 数据预处理       1. 数据合并       2. 重复值处理       3. 缺失值处理       4. 异常值处理    4. 本章小结 4. 特征工程    1. 特征构造       1. 语义特征构造       2. 元特征构造    2. 特征转换       1. 有序分类特征转换       2. 基于频率的特征转换       3. CatBoost目标编码       4. 数据标准化       5. TF-IDF向量化    3. 本章小结 5. 基础分类模型构建与训练    1. 模型构建       1. 逻辑回归模型       2. 决策树模型       3. 随机森林模型       4. 支持向量机模型       5. 神经网络模型    2. 模型参数设置与训练       1. 训练集与测试集划分       2. 模型训练    3. 模型评估       1. 评估指标选择       2. 各个基础分类模型评估结果分析    4. 本章小结 6. 多模型融合策略与实现    1. 多模型融合框架设计       1. 融合的目标与原则       2. 融合框架总体架构    2. 不同融合方法的实现       1. 投票法       2. 均值法       3. Stacking法       4. Blending法    3. 融合模型的评估与比较       1. 融合模型评估指标       2. 不同融合模型的性能对比    4. 本章小结 7. 模型部署与应用    1. 模型部署方案       1. 部署环境选择       2. 部署流程与步骤    2. 系统实现与测试       1. 在线招聘欺诈检测系统设计       2. 系统功能测试与验证    3. 实际应用效果分析       1. 应用场景与数据       2. 实际效果评估    4. 本章小结 8. 总结与展望    1. 研究成果总结       1. 主要研究内容回顾       2. 研究成果与创新点    2. 研究不足与展望       1. 研究过程中的不足       2. 未来研究方向及建议    3. 本章小结 9. 参考文献 10. 附录     1. 代码清单     2. 数据集详细说明及分析     3. 实验相关补充材料 11. 致谢 12. 论文的进度安排   2025年1月13日至2025年2月28日：阅读文献，熟悉研究现状，可以使用的研究方法、模型等；  2025年3月1日至2025年3月31日：尝试建立研究模型，整理研究思路；  2025年4月1日至2025年4月30日：优化模型，完成论文初稿；  2025年5月1日至2025年5月31日：修改论文，准备答辩资料 | | | |
| 指导教师审核：（包括选题大小是否合适、研究内容是否符合论文选题、是否符合本专业等等，上传时录入或者扫描件）  指导教师签名： | | | |